

Agrégation de méthodes statistiques pour la sélection de variables corrélées, en grande dimension

Aurélie Muller-Gueudin, Anne Gégout-Petit

► To cite this version:

Aurélie Muller-Gueudin, Anne Gégout-Petit. Agrégation de méthodes statistiques pour la sélection de variables corrélées, en grande dimension. JdS 2019 - 51emes Journées de Statistique de la SFDS, Jun 2019, Vandoeuvre-les-Nancy, France. hal-02360974

HAL Id: hal-02360974

<https://hal.archives-ouvertes.fr/hal-02360974>

Submitted on 13 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Agrégation de méthodes statistiques pour la sélection de variables corrélées, en grande dimension

Aurélie Muller-Gueudin
Anne Gégout-Petit

Université de Lorraine, Inria Nancy, équipe BIGS

JdS, 7 juin 2019

Sommaire

1 Contexte

2 Méthode

- Etape 1 : Prétraitement
- Etape 2 : Sélection de variables

3 Simulations

4 Données réelles

Problématique

Données transcriptomiques de patients atteints de cancer du poumon :

- $n = 37$ patients en stade avancé du cancer, ayant reçu une chimiothérapie.
- $m = 51\,336$ variables transcriptomiques, dépendantes entre elles.
- Une variable réponse Y : issue du traitement par chimiothérapie (binaire ou quantitative).

Objectifs

- 1 Sélection et tri des variables liées à Y
- 2 Visualisation des profils de patients

Sommaire

1 Contexte

2 Méthode

- Etape 1 : Prétraitement
- Etape 2 : Sélection de variables

3 Simulations

4 Données réelles

Prétraitement : structure de covariance des variables

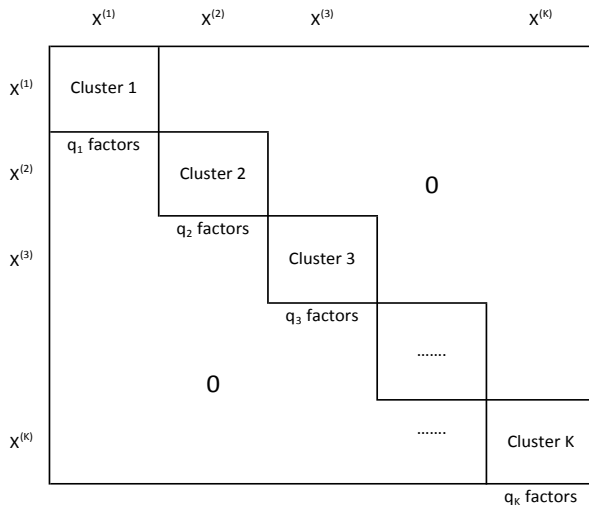


FIGURE 1 – Hypothèse sur la matrice de covariance des variables

Prétraitement : structure de covariance des variables



Hypothèses :

- $\mathbf{X} = (\underbrace{X_1^{(1)}, \dots, X_{m_1}^{(1)}}_{\mathbf{X}^{(1)}}, \dots, \underbrace{X_i^{(k)}, \dots}_{\mathbf{X}^{(k)}}, \dots, \underbrace{X_{m_K}^{(K)}}_{\mathbf{X}^{(K)}})$
- Indépendance des K clusters.
- Dans chaque cluster (k) , les $X_i^{(k)}$ sont liés à des facteurs $\mathbf{Z}^{(k)}$:

$$X_i^{(k)} = \mathbf{Z}^{(k)} \mathbf{b}_i^{(k)'} + \varepsilon_i^{(k)}$$

Prétraitement : structure de covariance des variables



Hypothèses :

- $\mathbf{X} = (\underbrace{X_1^{(1)}, \dots, X_{m_1}^{(1)}}_{\mathbf{X}^{(1)}}, \dots, \dots, \underbrace{X_i^{(k)}, \dots}_{\mathbf{X}^{(k)}}, \dots, \dots, \underbrace{X_{m_K}^{(K)}}_{\mathbf{X}^{(K)}})$
- Indépendance des K clusters.
- Dans chaque cluster (k) , les $X_i^{(k)}$ sont liés à des facteurs $\mathbf{Z}^{(k)}$:

$$X_i^{(k)} = f_i^{(k)}(Y) + \mathbf{Z}^{(k)} \mathbf{b}_i^{(k)'} + \varepsilon_i^{(k)} \quad (\text{Friguet 2010})$$

Détection des clusters

Algorithme ClustOfvar (Chavent et al, 2011)

Soit une partition $P_K = (C_1, \dots, C_K)$ de \mathbf{X} , on définit

- la variable synthétique s_k du cluster C_k :

$$s_k := \arg \max_{u \in \mathbb{R}^n} \left\{ \sum_{X_j \in C_k} r^2(u, X_j) \right\}$$

- l'homogénéité H du cluster C_k :

$$H(C_k) := \sum_{X_j \in C_k} r^2(s_k, X_j) = \underbrace{\lambda_1^k}_{1^{\text{e v.p. de ACP}}}$$

- l'homogénéité \mathcal{H} de la partition P_K :

$$\underbrace{\mathcal{H}(P_K)}_{\text{à maximiser}} := \sum_{k=1}^K H(C_k) = \sum_{k=1}^K \lambda_1^k$$

Synthèse de l'étape Prétraitement :

- 1 Détection de clusters, via ClustOfVar (Chavent et al 2011)
- 2 Décorrélation dans les clusters, via FAMT (Friguet, Causeur 2010)

A l'issue de cette étape, on note \mathbf{X}^* le jeu de données corrigé.

Exemple sur un jeu de données Peptides (800 variables)

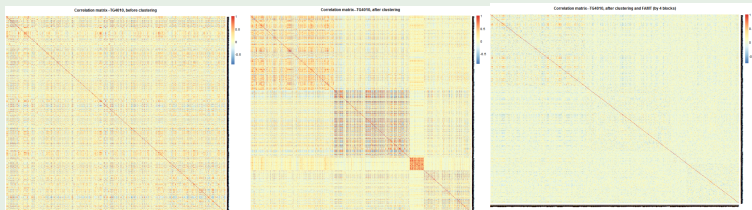


FIGURE 1 – Heatmaps des corrélations : avant clustering ; après clustering ; après FAMT par blocs.

Etape suivante : sélection de variables

Agrégation de méthodes statistiques

- On applique L méthodes statistiques de sélection sur (\mathbf{X}^*, Y)
- Chaque variable X_j obtient un score $S_j \in \{0, \dots, L\}$
où $S_j =$ nombre de sélections de X_j^* parmi les L méthodes.
 - ↪ Si $S_j = L$, la variable X_j influence +++ Y
 - ↪ Si $S_j = 0$, la variable X_j n'influence pas Y

On peut ensuite trier les variables selon leurs scores.

8 méthodes de sélection

- ❶ Bonferroni (1936)
- ❷ Benjamini & Hochberg (1995)
- ❸ q-value Storey & Tibshirani (2003)
- ❹ local FDR Aubert, Bar-Hen, Daudin, Robin (2005)
- ❺ FAMT Friguet & Causeur, 2010
- ❻ Forêts aléatoires : étape d'interprétation Genuer, Poggi, Tuleau-Malot (2010)
- ❼ Forêts aléatoires : étape de prédiction Genuer, Poggi, Tuleau-Malot (2010)
- ❽ Régression LASSO Friedman, Hastie, Tibshirani (2010)

⇒ Chaque variable X_j on a un score $S_j \in \{0, \dots, 8\}$: nb de sélections.

Sommaire

1 Contexte

2 Méthode

- Etape 1 : Prétraitement
- Etape 2 : Sélection de variables

3 Simulations

4 Données réelles

Simulations

- $n = 60$ observations
- Y binaire ($\frac{n}{2}$ observations avec $Y = 1$ et $\frac{n}{2}$ observations avec $Y = 0$)
- $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\dots,4} = 4$ clusters indépendants de 400 variables
- Variables corrélées (par cluster), via des facteurs latents \mathbf{Z} :

$$\mathbf{X}^{(k)} = \mathbf{f}_k(Y) + \mathbf{Z}^{(k)}\mathbf{B}^{(k)'} + \epsilon^{(k)}$$

Dans le 1er cluster :

400 variables = 40 variables influentes + 360 variables de bruit

$$X_j = \delta_j \mathbf{1}_{Y=0} + \mathbf{Z}b'_j + \epsilon_j$$

- $\delta_j = 1.5$ pour $j = 1, \dots, 10$
- $\delta_j = 1$ pour $j = 11, \dots, 20$
- $\delta_j = 0.75$ pour $j = 21, \dots, 30$
- $\delta_j = 0.5$ pour $j = 31, \dots, 40$
- $\delta_j = 0$ pour $j = 41, \dots, 400$.

Même schéma dans les 3 autres clusters.

⇒ Total : 160 variables influentes.

Intérêt de notre prétraitement

On compare 3 procédures de prétraitement

- ❶ rien n'est fait sur \mathbf{X}
- ❷ décorrélation de \mathbf{X} via FAMT :

$$\mathbf{X} \xrightarrow{\text{FAMT}} \mathbf{X}^*$$

- ❸ détection des 4 clusters via ClustOfVar, puis décorrélation dans chaque cluster via FAMT :

$$\mathbf{X} \xrightarrow{\text{ClustOfVar+FAMT}} \mathbf{X}^{**}$$

↪ Tests de Wilcoxon sur les 3 datasets issus de ces 3 procédures

↪ Sélection des variables X_j dont les p-valeurs sont $< 5\%$.

↪ 100 runs de (\mathbf{X}, Y)

Comparaison des 3 procédures

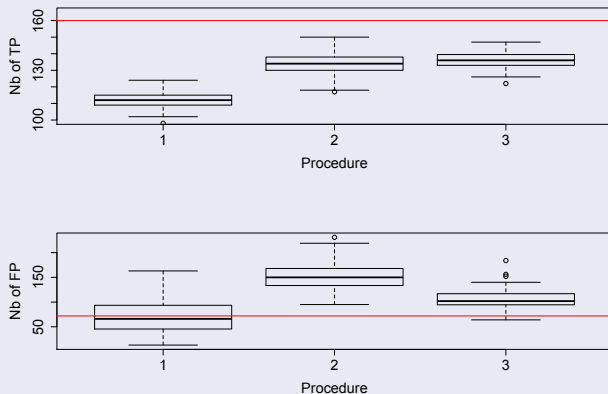


FIGURE 2 – Boxplots calculés sur $N = 100$ runs de (\mathbf{X}, \mathbf{Y}) . Lignes rouges : nombres attendus de TP (160) et FP ($5\% \times (1600 - 160)$)

↪ Amélioration du taux de FP, par rapport à FAMT.

Scores obtenus après prétraitement

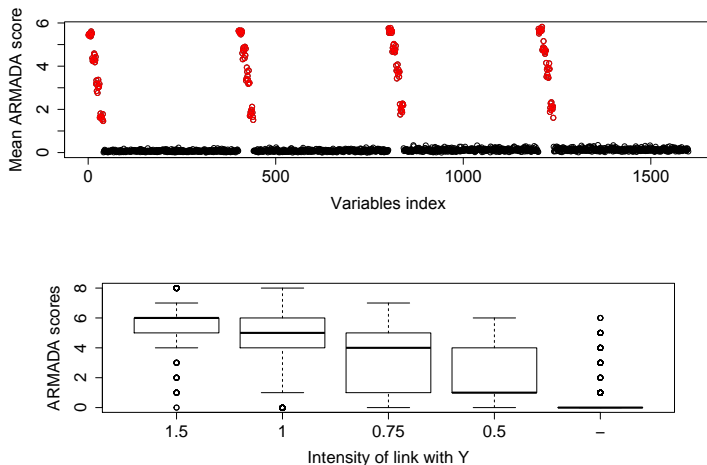


FIGURE 3 – Moyennes et boxplots des scores calculés sur $N = 100$ runs de (\mathbf{X}, Y) .

Comparaison avec d'autres méthodes de sélection

- armada : sélection de X_j si $S_j \geq 1$.
- Wilcoxon : sélection de X_j si $\text{pvaleur}(j) < \alpha$.
- FAMT : sélection de X_j si $\text{pvaleur ajustée}(j) < \alpha$.

	armada	Wilcoxon	FAMT
"1.5"	0.99 (0.04)	0.99 (0.07)	0.99 (0.02)
"1"	0.97 (0.15)	0.85 (0.35)	0.95 (0.20)
"0.75"	0.91 (0.27)	0.62 (0.48)	0.82 (0.38)
"0.5"	0.79 (0.40)	0.33 (0.47)	0.52 (0.49)
-	0.05 (0.23)	0.05 (0.22)	0.10 (0.30)

TABLE 1 – Taux moyens de sélection (avec écart-types) sur $N = 100$ runs de (\mathbf{X}, Y) .

↪ respect du taux de FP (5%) et meilleur taux de TP.

Comparaison de méthodes : courbes ROC

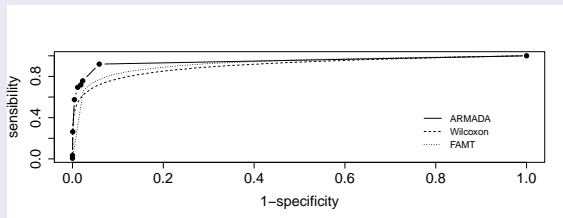


FIGURE 4 – Moyennes sur 100 courbes ROC obtenues sur 100 runs de (\mathbf{X}, Y) .

↪ Courbe ROC armada au dessus des autres.

Remarque : mêmes résultats sur des simulations en régression.

Sommaire

1 Contexte

2 Méthode

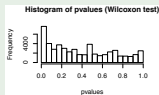
- Etape 1 : Prétraitement
- Etape 2 : Sélection de variables

3 Simulations

4 Données réelles

Données transcriptomiques, patients traités par chimio

- $n = 37$ patients (13 décédés à 12 mois + 24 vivants à 12 mois)
- $m = 51\,336$ variables



- Filtre des variables avec Wilcoxon-pvalue $> 5\%$
 $\implies m = 6810$ variables

- 1 Etude en classification : $Y = 1/0$ (pour "décédé" / "vivant" à 12 mois après la chimio)
- 2 Etude en régression : $Y =$ temps de survie après la chimio (pas de censure)

	Classification score							
Regression score	0	1	2	3	4	5	6	7
0	2227	328	273	337	531	257	34	1
1	41	7	3	9	17	10	2	0
2	131	35	39	52	119	71	9	0
3	119	48	44	50	117	114	17	0
4	174	65	56	86	256	241	102	4
5	119	64	40	57	116	176	116	4
6	15	4	4	5	12	19	26	1
7	1	2	1	0	1	0	0	0
8	0	0	0	0	1	0	0	0

TABLE 2 – Répartition des scores des variables. 342 variables ont des scores ≥ 5 en classification et en régression.

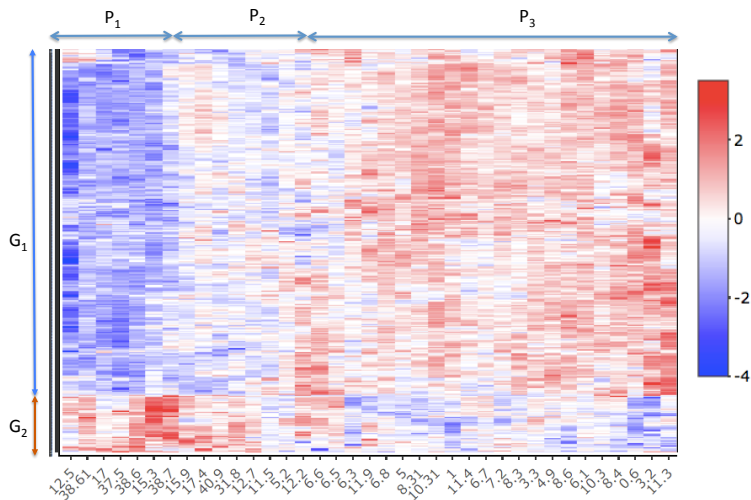


FIGURE 5 – Heatmap des 342 variables ayant des scores ≥ 5 en classification **et** en régression. 1 colonne = 1 patient (marqué par son temps de survie), 1 ligne = 1 variable.

Robustesse des résultats

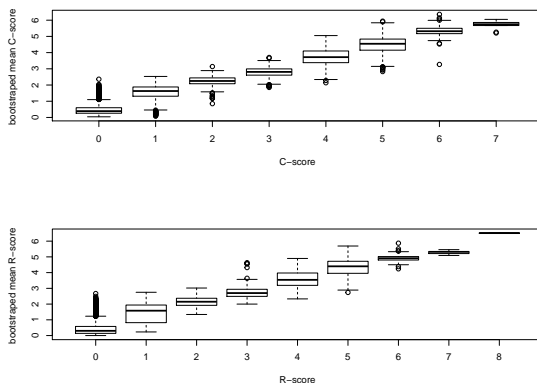


FIGURE 6 – Distribution des moyennes des C-scores et R-scores obtenues sur $B = 100$ échantillons bootstrap, versus les scores originaux, pour toutes les $m = 6810$ variables.

Conclusion

- Sélection de variables (ici gènes) liées à une variable d'intérêt (ici issue d'un traitement)
- Variable d'intérêt binaire, multinomiale, ou continue
- Visualisation, sur tous les patients, des gènes sélectionnés
- Classification de profils génétiques de patients
- \implies développement de la médecine personnalisée....
- A été utilisé par des biologistes pour comprendre la fonction des gènes liés à ER36 (article en cours)

Package *armada* disponible sur le CRAN.